DOCUMENT RESUME

ED 294 915 TM 011 635

AUTHOR Loewen, James W.; And Others
TITLE Gender Bias in SAT Items.
SPONS AGENCY Women's Educational Equity Act Program (ED), Washington, DC.
PUB DATE Apr 88
NOTE 63p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Academic Ability; Grade 10; Grade 11; Grade 12; High Schools; *High School Students; Item Analysis; *Sex Bias; Sex Differences; Test Anxiety; *Test Bias; Test Coaching; Test Construction; Test Items
IDENTIFIERS *Scholastic Aptitude Test

ABSTRACT

Sex-related bias on the Scholastic Aptitude Test (SAT) was studied in a sample of 1,112 students in SAT coaching classes who took the SAT. Of these, 1,028 answered an additional questionnaire (Appendix A of this report) about high school grade point average, perceived abilities, and background. Almost all of the subjects were 11th graders (97.8%), with 0.7% sophomores and 1.5% seniors. Seven items on the verbal and 10 on the math section showed more than a 10% difference in the percentage of each sex getting them correct. An additional 22 verbal items and 16 math items showed a difference greater than 5% in favoring one sex. Other questions considered were methods of test construction that might reduce bias, the relationships between SAT scores and school performance, the role of test anxiety on SAT performance, and the effects of SAT scores on college choices. As other researchers have found, girls did less well than boys on the SAT, yet had higher school grades. Recommendations for sex-fairer tests include the following: (1) remove items having large response differences between the sexes unless balanced by other items; (2) manipulate mean scores on the verbal test so that males and females score equally; (3) ensure that verbal content on math items favors girls; and (4) make validity studies correlating test scores and first-year college performance widely available both to consumers and researchers. Nineteen tables and two graphs are included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED294915

"Gender Bias in SAT Items"

James W. Loewen, Department of Sociology,
University of Vermont, Burlington, VT 05405,

Phyllis Rosser, Equality in Testing Project,
77 Telegraph Hill Road, Holmdel, NJ 07733, and

John Katzman, Princeton Review,
606 Columbus Avenue, NY, NY 10024.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

JAMES W. LOEWEN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

American Educational Research Association
New Orleans, LA
April 5, 1988

For further information or networking, please contact the Equality in Testing Project.

The activity which resulted in this report was supported by a grant from the U.S. Dept. of Education, under the auspices of the Women's Educational Equity Act. Opinions expressed herein do not necessarily reflect the position or policy of the Department, and no official endorsement should be inferred.

Discrimination Prohibited: No person in the United States shall, on the grounds of race, color, or national origin, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any program or activity receiving Federal financial assistance, or be so treated on the basis of sex under most education programs or activities receiving Federal assistance.

© Copyright 1988 by James W. Loewen, Phyllis Rosser, John Katzman

TM 011 635

I. Introduction.

In 1987 women averaged 57 points lower than men on the Scholastic Aptitude Test (SAT): 47 points lower on the math section and 10 points lower on the verbal. The purpose of the SAT is to predict first-year college grades, but numerous studies show that women receive higher average grades in all subjects, in both high school and college classes, yet receive lower average SAT scores. (Clark and Grandy, 1984; Ramist and Arbeiter, 1985; Cordes, 1986) The test publisher, Educational Testing Service (ETS), admits that the SAT "slightly under-predicts college grades for females and over-predicts college grades for males." (Clark and Grandy, 1984, p.20)

Years ago, men's higher math scores were partly offset by women's higher verbal scores. In 1967, for instance, women averaged the same 47 points lower on the math section, but 5 points higher on the verbal. (College Entrance Examination Board [CEEB], 1987) However, females lost their verbal lead in 1972, as a result of gradual changes in item content to include more items referring to science, business, and "practical affairs," and fewer to human relations, arts, and humanities (Dwyer, 1976a). The test was changed to create "a better balance for the scores between the sexes" (Donlon and Angloff, 1971, pp. 25-26), and by 1986, men's verbal scores averaged 11 points higher than women's. Dwyer noted that test specifications have been changed to mandate more male-oriented items on verbal tests, where females traditionally excel, but the reverse (more female-oriented items

on math tests, where males traditionally excel) has not been mandated; she calls this "nonconscious sexism" (1976b).

The verbal advantage that males now enjoy on this test and its companion, the Preliminary Scholastic Aptitude Test (PSAT) is unusual. Women now also score less than one point lower than men on the GRE-Verbal. Otherwise, we have surveyed score results from a variety of standardized aptitude and achievement tests, and on all but one, women receive higher verbal scores. In light of the research literature, which "either finds no difference between men and women in cognitive skills or finds a slight advantage for females on verbal skills and a slight advantage for males on mathematical and spatial skills" (Clark and Grandy, 1984, p.4), the male SAT and PSAT advantage is extremely questionable.

At the same time, it is extremely important, for it directly influences National Merit Scholarship awards, 2/3 of which go to boys, state merit scholarships, admissions to some colleges, and participation in programs for gifted high school students (Rosser, 1987). To investigate this male advantage, we conducted an item analysis of one form of the June, 1986, SAT. We wanted to determine whether specific questions were creating or widening the score gap between the sexes, to investigate other factors that might contribute to sex differences in SAT scores, and to see how SAT scores influenced students' future academic plans.

Our purposes included:

--to learn what test items, if any, showed marked sex-related (gender-related) biases, favoring girls or boys;

--to investigate item-to-scale (point-biserial) correlations of sex-biased items, in order to study methods of test construction that might reduce sex bias;

--to investigate relationships among SAT scores, high school grade point averages (GPAs), and sex, to see if girls' lower SAT scores were accompanied by correspondingly lower school performance;

--to investigate other factors, such as socioeconomic status (SES), test anxiety, and high school subject preference, that might help explain why girls do worse than boys on the SAT but not in high school or college; and

--to investigate effects of SAT scores on students' college choices and self-perceived abilities, by sex.

II. Methods.

In March, 1987, 1112 students in Princeton Review coaching classes took a form of the June 1986 SAT, during the second session of their coaching classes, under conditions as similar as possible to those in ETS test centers. As the final section of the exam, 1028 students answered our additional 25-item questionnaire (Appendix A), which asked them to indicate their high school grade point averages (GPA), favorite high-school

subjects, perceived ability in English and math, test anxiety, family background, etc.[1]

Sample: Because not every student answered every item, our N ranged from 1112 on some SAT items to about 1010 on some questionnaire items.[2] They came from the five boroughs of New York City, from public schools like Bronx Science and Stuyvesant, nonselective public schools, parochial schools, and private schools such as Dalton. They were fairly closely balanced between the sexes, 55.6% girls, 44.4% boys. (Nationally, SAT takers are 52% female.) They were mostly white (75.3%), but included 13.2% Asian-Americans, 5.2% blacks, 2.4% Hispanic, and 3.9% other and blank. Almost all (97.8%) were in 11th grade; 0.7% were sophomores, and 1.5% were seniors.

Students' high school preparations were rather strong. In self-reported high school grade point average (GPA), 57% reported averages from B+ to A+. Nor were these grades earned in easy courses. In math, including their current (junior) year's classes, 86% had taken three years, one course per year; 11.7% had taken more. English preparation was also strong. 92% of our girls and 91% of our boys had taken 3 years of English in their 3 years of high school. About 7% had taken more. In natural

---

[1]ETS also relies on self-reported data for its analyses; studies have found rather high correlations (.7 to .9) between self-reports and corresponding objective measures (Clark and Grandy, 1984).

[2]Owing to slightly different procedures for determining sex of student on different computer runs, N's and %'s can vary slightly from table to table.

science, 73% of all students had taken three years; among the rest, boys were more likely to have taken an additional year while girls were more likely to have taken less.

Our students came mainly from upper-middle class backgrounds. 81% of their fathers and 52% of their mothers had professional careers (doctors, executives, engineers, teachers, etc.). 72% of their fathers and 63% of their mothers were college graduates. 60% of our sample attended public school, 9.5% parochial, and 27% prep school, with little difference between the sexes, except that 4% more males were attending parochial schools.

Because our sample came from one metropolitan area and selected themselves by paying for an expensive coaching course, they cannot be seen as random or representative of the national population. However, we think it offers a valid way of making internal comparisons -- boys versus girls, anxious versus not anxious, etc. -- and we believe that the processes we found operating within this sample can probably be generalized to others. The uniformity of our sample regarding SES allows us to explore differences by sex that cannot be attributed to low incomes or less educational preparation.

III. Results: Sex Differences.

The SAT is scored on a 200-800 scale. It consists of 85 verbal items, each worth about 7 score points, and 60 math items,

each worth about 9.5 score points.[3] On the verbal SAT nationally, boys now outscore girls by about 10 points, or 1.4 items; our boys and girls were about equal. On the math scale, boys outscore girls nationally by about 47 points or 5 items;

Table 1. SAT Averages by Sex.

Group	Verbal		Math		Total
	Raw	Scale	Raw	Scale	Scale
Female, National		425		453	878
Male, National		435		500	935
Female, Our Sample	44.8	489	33.6	536	1025
Male, Our Sample	45.0	490	37.1	571	1061

among our students, boys outscored girls by 3.5 items or about 35 scale points.

[3]We did not examine the TSWE (Test of Standard Written English). When computing our sample's math scores, we assumed all math items had 5 alternatives. Some have 4, so our procedure slightly under-subtracts for wrong answers, giving our students slightly (<10 points) higher math scores than they should have. ETS uses an irregular scale to convert raw scores to SAT scale scores; hence getting one more item correct can increase SAT scores by 0, 10, or 20 points. We converted their scale to regular intervals.

SAT Items Showing Major Sex Differences.

Girls and boys scored within a few percentage points on most verbal and math items, reflecting the facts that wide areas of experience, skills, and sub-cultural terms are shared by young people of both sexes, and that most SAT questions tap those areas. However, we were surprised to find that 7 items on the verbal and 10 on the math sections of the SAT showed considerable (>10%) differences in % of each sex getting them correct. Table 2 lists the verbal items, with those favoring girls indicated by a + sign.

Table 2. 7 SAT Verbal Items Favored One Sex by Approximately 10% or More.

Section, Item #, Description	Female % - Male %
1 #1, "setback," opposite "improvement"	-10.7%
1 #5, "sheen," opposite "dull finish"	+18.3
1 #23, author's tone, science passage	-11.8
1 #44, "mercenary is to soldier"	-15.7
4 #21, "pendant is to jewelry"	+9.6
4 #24, "love is to requite"	+14.5
4 #31, "betrayal"	+10.2

It is not surprising that words referring to relationships ("requite"), jewelry, and fabric ("sheen") favor girls; conversely, "mercenary" relating to "soldier" is a male-loaded term in a society that drafts only men for military service. Previous studies (Coffman, 1961; Strassberg-Rosenberg and Donlon, 1975; Dwyer, 1979) have found that item content produces important sex differences in performance. Recent public concern with item bias and wording has led ETS to create a Sensitivity Review

Process (ETS, 1987) for all items. However, Table 2 indicates that the Sensitivity Review does not eliminate items that favor one group. In Part IX of this paper we suggest problems with the ETS approach.

Among the 85 verbal items, 22 additional items favored one sex or the other by >5%, a cutoff point suggested by Green (1987).

Among math items, 10 differences of greater than 10% appeared, all favoring men (hence differences are marked -).

Table 3. 10 SAT Math Items Favored One Sex by >10%.

Section, Item #, Description	Female % - Male %
2 #8, "liters per hour"	-10.3%
2 #15, "chore 994th boy will have at boys camp,"	-12.3
2 #16, "number of boy with chore at boys camp"	-15.6
2 #19, "parallelogram ratios"	-12.2
2 #20, "1/6 as decimal, sum of digits"	-10.7
2 #21, "basketball team won/loss record"	-27.0
2 #22, "<(a-b)<"	-11.0
2 #25, "n as odd integer"	-10.8
5 #17, "length of right triangle"	-10.7
5 #25, "inequalities with X^2, -x"	-10.6

Three math items -- #15, 16, and 21 -- were about boys' enterprises, suggesting that verbal bias adversely affects girls' performance on math items.[4] Earlier studies have shown that when math content is made relevant to female experience, males do not

[4]Although high-school basketball teams may now be female as well as male, we think the average reader infers "male" when the context is a concern about won/loss record. Moreover, figuring these kinds of statistics is a common activity among adolescent males.

outperform females on math problems (Milton, 1958; Bem, 1970; Graf and Riddell, 1972; Donlon, 1973; McCarthy, 1975). Items on which males markedly outperformed females ranged in difficulty from easy to hard, implying that the level of mathematics involved did not cause the difference in performance by sex.

Among the 60 math items, 16 additional items favored one sex or the other by >5%.

Students in the Middle Range Were Most Affected by Sex-Biased Items.

We suspected that middle-range scorers would be most affected by sex-biased items. High scorers might be more likely to be certain of the right answer, while low scorers might not know anything about the right answer, so they might not even guess at it. Middle scorers might know just enough to guess, but their "subliminal" knowledge would be more easily misled by sex-biased items. Table 4 divides our sample into 4 groups by overall verbal scores (low = 200-480, low-middle = 481-530, high-middle = 531-580, and high = 581-800), and again by math scores (200-520, 521-580, 581-650, and 651-800).[5] For all items listed in the previous to tables, Table 4 displays the mean absolute differences (% of males getting the items correct minus % of

[5]As the overall mean differences imply, girls and boys group about the same in verbal scores, while 11% more girls end up in the low math group, compared to boys. Thus the overall male/female difference in math exam performance is greater than Table 4 displays, because more males than females fall in the higher columns of the table.

females, for male-biased items; the reverse for female-biased items). Middle scorers were affected most, though the differences were small.

Table 4. Mean Differences by Sex in Percentage Correct on Sex-Biased Items, Among Low, Middle, and High SAT Scorers.

	SAT Score Range				
	Low	Low-Mid	High-Mid	High	All
Among Verbal Items:	14.3	14.1	16.5	7.9	13.0
Among Math Items:	4.6	6.0	7.9	6.9	-12.0

Do Certain Forms of Items Favor One Sex?

ETS divides verbal items into four types: antonyms, reading comprehension questions, sentence completions, and analogies. Contrary to studies that found women Strassberg-Rosenberg and Donlon, 1975) and blacks (Schmitt and Dorans, 1987) doing better on reading comprehension items and worse on analogies (Donlon, 1973; Stricker, 1982), we found that girls and boys performed about the same on all item types. Girls were better at antonyms, worse at reading comprehension, but the differences were slight. We also found no important differences by difficulty of item on the verbal test.

We classified math items into four types: computation, geometry, algebra, and problem solving. Prior research has been equivocal as to which sex does relatively better on which types. Donlon (1973) found that females performed relatively better in

algebra than geometry, while Milton (1957) and Graf and Riddell (1972) found that problem solving favored boys. Becker (1983)

Table 5. Scores by Sex on Different Types of Items.

Type of Questions	Average Percentage Correct		
	Female	Male	Female % - Male %
Antonyms	62.2%	60.9%	1.3%
Reading Comprehension	46.5	47.8	-1.3
Sentence Completion	71.0	71.5	-.5
Analogies	66.0	65.6	.4
10 Easy Verbal Items	87.5	88.5	-1.0
10 Medium Verbal Items	59.6	57.9	1.7
10 Difficult Verbal Items	25.4	25.4	0.0
26 Algebra Items	62.5	67.8	-5.3
14 Geometry Items	54.1	58.8	-4.7
14 Computation Items	71.5	74.9	-3.4
6 Word Problems	60.6	65.8	-5.2
10 Easy Math Items	85.9	86.4	-0.5
10 Medium Math Items	55.7	63.2	-7.5
10 Difficult Math Items	19.2	28.0	-8.8

found SAT algebra items more difficult for junior high girls than boys, but no sex differences in geometry and computation. McPeek and Wild (1987) found women performing better on algebra than geometry on the GRE. We found nothing to substantiate consistent sex differences. Girls scored closer to boys on computation, but the difference was slight; they were no better on algebra compared to other math areas. Nor did we find important differences by difficulty of item; boys outscored girls on all

but 8 items, although the differences were predictably smallest on the easiest items.

Table 5 indicates that type of item differentiated between males and females much less than item content, as shown in tables 2 and 3.

IV. Do SAT Sex Differences Correlate With Performance Differences?

Since we could not correlate SAT scores with first-year college grades, we used a surrogate: high school GPA. Researchers have consistently found that high-school GPA is the best single predictor of college GPA, and although its r is only about .48, that is higher than r's for the SAT or most other predictors (ACT, 1973; etc.).

In our sample, SAT scores correlated only moderately with high school GPA: r between the V-SAT and High school GPA = .28, while r between the M-SAT and high school GPA = .33. These r's are similar to the r's of .3 between SAT scores and first-year college grades reported by Schrader (1984), but lower than the r's of .5 between SAT scores and high school rank in class in Schrader's national study.

Girls in our sample are performing considerably better in high school than their relative SAT scores would suggest. Although they received lower scores than boys on both parts of

the SAT, Table 6 shows they are getting better grades. Thus this SAT is under-predicting girls' high school GPAs.

Table 6. Percentage Reporting Various High School GPA's, by Sex.

GPA	Percentage of Girls	Percentage of Boys
A to A+	18.7%	15.3%
B+ to A-	43.0	36.8
B- to B	30.8	39.8
C+ or Lower	6.0	7.4

Another way of showing this under-prediction is to try to use SAT scores to predict high school GPA's, by sex. Table 7 shows that the SAT "predicts" high school GPA well, within each sex. But note the female/male differences. Within almost every SAT score category, looking across the top data row, a higher % of girls get A to A+ grades than boys. For example, 41.7% of

Table 7. Percentage Reporting Various High School GPA's, by SAT Score Range and Sex.

% Receiving GPA of:	Percentage of Girls With Verbal SATs				Percentage of Boys With Verbal SATs			
	Low	Low-Med	Med-High	High	Low	Low-Med	Med-High	High
A to A+	4.2%	14.3%	24.1%	34.8%	4.4%	12.9%	17.7%	27.9%
B+ to A-	37.6	41.4	43.6	50.4	24.8	34.7	40.6	49.2
B- to B	43.0	36.1	29.3	12.8	56.9	43.6	37.5	19.7
C+ or Lower	12.7	6.8	2.3	0.7	13.8	7.9	2.1	3.3
% Receiving GPA of:	Percentage of Girls With Math SATs				Percentage of Boys With Math SATs			
	Low	Low-Med	Med-High	High	Low	Low-Med	Med-High	High
A to A+	3.1%	12.0%	25.3%	41.7%	2.6%	4.7%	14.5%	31.4%
B+ to A-	30.6	47.3	51.3	43.5	14.1	28.0	48.1	45.7
B- to B	49.4	36.0	19.5	12.0	60.3	59.8	31.3	21.4
C+ or Lower	14.4	4.7	1.9	0.9	23.0	6.5	5.3	0.7

girls with top Math SATs get A to A+ grades, while only 31.4% of boys do. This trend even continues for B+ to A- grades, which is surprising since some girls get "used up" in the top row.

We also compared grades in high school English courses with VSAT scores, and grades in high school math courses with MSAT scores. Again, controlling for SAT scores, more girls got A to A+ grades, compared to boys, in English and math. These findings agree with CEEB validity studies cited by Clark and Grandy (1984) that show women receiving college grades equal to or better than men's in math, science, and the humanities. Massachusetts Institute of Technology has also found that women with lower SAT math scores earn college grades equal to men and has changed its

admissions policies accordingly (Behnke, 1987).

All this raises an important question about the SAT. If women score worse than men, yet earn better grades in high school and college, shouldn't it be changed to eliminate female underprediction? In Part IX we make policy recommendations on this point. First, we assess various factors that have been suggested to account for the female underprediction.

V. Student Factors That May Cause Sex Differences in SAT Scores.

Test Anxiety

Researchers have suggested that test anxiety may create different performance by sex on the SAT. Indeed, our females reported considerably more test anxiety, as Table 8 shows.

Table 8. "How Do You Feel About the SAT?" by Sex.

Level of Anxiety	Girls	Boys
"extremely anxious"	27.8%	10.8%
"moderately anxious"	38.5	37.7
"somewhat anxious"	24.9	34.2
"not anxious at all"	8.8	17.3

There were 2 1/2 times as many "extremely anxious" girls as boys. Girls' anxiety may constitute a rational response to their history of lower SAT performance, compared to their high school grades. However, in our sample, test anxiety did not correlate closely with poor test performance, particularly among boys.

Among girls, the least anxious group scored considerably worse than others. "Extremely anxious" girls scored lower on the math SAT than "somewhat" anxious, but anxiety levels had no effect on verbal scores.[6]

High school GPA also had no systematic relationship with test anxiety. Nor did students' own ratings of their verbal and math skills. However, we did find that the more anxious the test taker, the more likely s/he was to claim that tests underrate their abilities. SES also influenced anxiety: students whose fathers were professionals were less anxious than those whose fathers were not professionals. Mothers' occupation made no difference to sons, but 33.7% of daughters of women who don't work outside the home were "extremely anxious" about the SAT, compared to 23.1% of daughters of mothers with professional careers.

Parents' education had mixed impact, but generally, children of more educated parents were less anxious. Anxiety correlated moderately with plans to attend "super-elite" colleges. Because test anxiety is unpleasant and unproductive, we recommend research to investigate further why girls and upwardly mobile

---

[6]Of all issues we studied, anxiety versus performance was probably most affected by our test conditions. Students' performance on this SAT did not "count." On an SAT upon which college entrance and scholarships depended, anxiety might hurt performance more. Also, boys may not have admitted as much test anxiety as girls to us, but may actually feel as anxious.

Faigel (AP, 1987) found that students with unusually high test anxiety performed poorly; after taking a drug used to treat high blood pressure, their verbal scores rose by 50 points and their math scores by 70.

boys are more anxious, so that steps can be taken to decrease their anxiety. Of course, so long as students perceive that their educational futures are at stake and that they test below their true ability, test anxiety may not be ameliorable by programs based on research.

Time Pressure.

Graf and Riddell (1972) found that on math problems perceived to be more difficult, girls proceeded more slowly than boys. Others have found no appreciable sex differences in test-taking speed (Donlon, 1977; Wild, Durso, and Rubin, 1982). To determine whether either sex was more affected by time pressure, we examined performance on the last 10 items on the last two tests, section 4 (verbal) and 5 (math), and found no important differences by sex. Girls did slightly better than boys on the final verbal items; boys did better on the final math items, just as they did on earlier math items. Almost identical percentages of boys and girls left the last 5 items blank on the verbal test; on the math test, 5.5% more girls left them blank, but slightly more girls than boys left earlier math items blank as well.

Liking Mathematics Helps Math SAT Scores.

To explain the large gender gap in math scores, researchers have suggested that prior socialization influences boys to like math more than girls and to take more math courses in high school. We investigated these factors in our sample.

36.2% of our boys chose math as their favorite subject or science first and math second, compared to 22.4% of our girls. Another 13.2% of boys and 11.6% of girls chose math as their second favorite subject. Liking math raised scores on the math SAT for both sexes, as Table 9 shows. The male/female gap remained, though it narrowed somewhat. Among math-likers, for instance, males held a 2.6 point advantage, while in total scores, all males had been 3.5 points ahead. Of course, liking math may also partly be a result of good scores on prior "standardized" tests. Liking math may also correlate with taking math, which we will investigate in the next section.

Table 9. Math SAT Items Correct by Math as Favorite Subject, by Sex.

Mean # Items Correct	Math First	Math Second	Math Not Chosen
Among girls:	40.9	38.5	36.0
Among boys:	43.5	41.4	37.7

Interestingly, students of both sexes who chose math as their favorite subject earned lower scores on the verbal SAT. This is understandable if we believe that students who like math don't like English, but it is not understandable if we believe that math is a difficult subject, likers of which might be more studious, hence better in all subjects. We also found several items on which math-likers did much (>10%) worse than math-dislikers; only one favored math-likers by 10%. This subject warrants further study.

Taking Mathematics.

ETS notes that the math SAT doesn't utilize math beyond algebra and geometry, so students who had taken more advanced math shouldn't be advantaged by that fact. (Of course, such students might like math or have demonstrated math skills to get into such courses.) Table 10 shows that most of our students, regardless of sex, had taken one year of math per year in school. Only 23 students had omitted a year or more of math. Although 18 of these were girls, the percentage of all girls taking less than the typical three years of math was only 2.9%. 15.7% of our boys

Table 10. Math Preparation by Sex.

Years of Math in High school	% of Students		
	All	Girls	Boys
4 or more	11.7%	8.6%	15.7%
3 (one/year)	86.1	88.5	82.8
2 or less	2.2	2.9	1.5

(Not adjusted for the 2.5% non-juniors.)

took more than 3 years of math, compared to 8.6% of our girls, fairly similar to national studies (Ramist and Arbeiter, 1985). However, extra math did not affect SAT performance much, probably because higher math is not required for SAT math questions. Table 10 primarily indicates that most of our sample have had one year of math in each year of high school, regardless of gender.

Table 11 shows the relationship of years of math to SAT scores. Unlike "likes math," "takes math" does not adversely

affect verbal SAT scores, while it correlates with higher math SAT scores. Controlling for years of math taken slightly narrows the F/M gap in Math SAT scores: the largest group in our sample, students in the "3 Years" column, show a 2-point gap, a bit less than the 3.5-point gap in the entire sample. Girls taking less math than average exhibit a 5-point deficiency compared to boys. Like a National Assessment of Education Progress (NAEP) study (Welch, Anderson, and Harris, 1982), our males still did somewhat better after the effects of differential preparation were removed.

Table 11. Math SAT Items Correct by Amount of Math Taken, by Sex.

Mean # Items Correct	4 Years	3 Years	<3 Years
Among girls:			
Verbal SAT	52.6	51.0	49.4
Math SAT	42.1	37.6	33.9
Among boys:			
Verbal SAT	51.6	50.7	51.9
Math SAT	43.8	39.5	39.0

Interestingly, taking math helps performance on some verbal items, while hurting performance on others. Moreover, on some items, taking math helps girls but hurts boys, while on others, taking math hurts girls but helps boys! Further research into these items may help explicate test-taking styles of men and women.

VI. SES Factors That May Cause Sex Differences in SAT scores.

Parental Education.

Like other researchers of SAT performance, we found that social class, measured by fathers' and mothers' educations and occupations, had immense impact on scores.[7] Daughters and sons of more educated fathers (more than BA) averaged about 8.5 more verbal SAT items correct compared to children of less educated fathers (no college). Daughters of educated fathers did better on all but 6 of the 85 items, and >10% better on 39 than daughters of less educated fathers. Sons of educated fathers did better on all but 5 of the 85 items, compared to sons of less educated fathers, and >10% better on 48! In math, daughters of more educated fathers averaged 5.3 more items correct than daughters of less educated fathers; sons varied by 10.3 items. Daughters of educated fathers did better on all but 3 of the 60 math items, and >10% better on 20. Sons of educated fathers did better on all but 9 of the 60, and on those 9 did about the same, while they did >10% better on 24 items.

Mothers' education affected boys even more than girls on the verbal test: sons of more educated mothers got 11.5 more verbal items correct, compared to sons of less educated mothers, while daughters varied by 8.6 items. Daughters of more educated

---

[7] A weakness of our analysis is that we did not investigate whether these students lived with their fathers, mothers, or both.

mothers did better on all but 2 verbal SAT items compared to daughters of less educated mothers, and they did >10% better on 39. Sons of more educated mothers did better on every verbal SAT item, and they did >10% better on 55! On math items, children of more educated mothers got 5.7 more math items correct than children of less educated mothers. Daughters of more educated mothers did better on all but 1 math SAT item compared to those of less educated mothers, and they did >10% better on 28 of the 60 items. Sons of more educated mothers did better on all but 3 math SAT items compared to those of less educated mothers; on 27 items, the difference was greater than 10%.

Parental Occupation.

Fathers' occupations influenced SAT scores. This is not surprising: for decades, ETS has reported high positive correlations between parental income and SAT scores. Fathers' occupations made about twice as much difference for boys as for girls.

Mothers' occupations made the same kind of difference as fathers', as Table 12 shows; children of professionals had higher scores than children of mothers with "other" occupations. Mother's occupations hold additional interest owing to the category "works in home," which differs from other occupations in that it is not itself related to social class. It is more traditional, however, compared to mothers who work outside the home. And it made a big difference to scores, especially among

girls. Table 12 indicates that not working outside the home had about the same effect as holding "other" occupations (lower middle class, such as real estate, social worker, and upper working class, such as sales clerk, waitress). Yet many such mothers are married to high professional husbands (as are some "other" occupation mothers). Thus social class itself probably did not underlie this difference.

Table 12. Mean # of SAT Items Correct As Affected by Mother's Occupation, by Sex.

	Among Students With Mother's Occupations:			
	Professional	Other	Works in Home	Diff.(Col.1-3)
Girls, Verbal	54.5	48.2	48.8	5.7
Boys, Verbal	55.3	50.1	48.7	6.6
Girls, Math	38.1	36.1	36.7	1.4
Boys, Math	42.8	38.7	39.9	2.9

Girls whose mothers worked only in the home perceived their English ability to be lower than girls whose mothers had professional careers, and scored lower on the verbal SAT, although their high school grades in English were equal. This suggests a link between SAT scores and self-esteem. Among boys, mothers' occupation did not affect perceived English ability. Regarding perceived math ability, the picture reversed: mothers' occupation made little difference to daughters, but did relate positively to sons' perceptions and to sons' math GPAs.

Perhaps mothers who work at home have lower self esteem which they may pass on to their children, a possibility suggested by the research of Jacobs and Eccles (1985) on math ability; further studies are required to substantiate this possibility.

Although not central to our focus on gender differences, we cannot leave this section without pausing to emphasize how pervasive the influence of social class was on our students' scores. This is all the more striking in view of the constricted social class range among our students' families.

Of course, class influences on SATs have been pointed out many times before; class also underlies some (although not all) of the gap between black and white scores. We would stress that on some items, higher SES students scored >10% above others; indeed, on several items, they scored >20% better. On other items, SES made little difference. We suggest that some items are probably "classist," the same way some have proven in this study to be sexist, and the suggestions we make later for strengthening ETS's item-selection process to promote gender fairness would hold even more strongly regarding class and race.

VII. Effects of the SAT on Students.

Our sample had a good self-image regarding their own abilities. In "reading and writing ability," a majority (57.3%) placed themselves in the top 10% of their peers, while only 1.7% were in the bottom half! (Girls and boys were almost identically positive.) In math ability, girls were less sure: 38% of them but 56% of boys claimed to be in the top 10%.

Our students also showed healthy self-images or serious criticism of "standardized" tests in response to the question, "Do you feel your past test scores on standardized tests (PSAT, etc.) are accurate?" 81.3% claimed their "ability is higher than the tests indicate." There were no important sex differences.

SAT Differences, High school GPA Differences, and Perceived Ability, by Sex.

"Standardized" tests can hurt students by adversely affecting their self-image. Students with bad test scores may reasonably infer that they have low "verbal aptitude" or "math aptitude," since ETS uses "aptitude" to title their tests. Almost all of our students had taken ETS tests previously, and their scores on this SAT can be taken as a surrogate for prior scores. We then compared test feedback to teachers' feedback (high school grades), to see which had greatest impact on students' own reports of their verbal and math abilities. Tables 13 and 14 show ratios of girls' scores to boys'. When girls and boys were equal, the ratio is 1; if girls scored better, the ratio is >1; if girls scored worse, the ratio is <1.

Table 13. Ratio of Girl/Boy Ranking on English HS GPA, Verbal SAT Scores, and Perceived Verbal Abilities.

Item	Girl Result Divided by Boy Result
% A+ on English HS GPA	1.58
% A- through A+ on English HS GPA	1.20
% in Highest Group on Verbal SAT	.95
% in Highest Two Groups on Verbal SAT	1.01
% Estimating Their Verbal Ability in Top 5%	1.11
% Estimating Their Verbal Ability in Top 10%	1.05

Girls did better in English in school than boys, but about the same on the verbal SAT. They ranked their English abilities only a little higher than boys, in line with the SAT results. Inmath, girls did about as well in school, but worse on the math SAT. Again, they estimated their math ability in line with the test results, not the classroom results. Thus, although girls and boys got almost identical grades in math, only 38% of girls put themselves in the top 10% in math ability, compared to 56% of boys. Therefore, like Clark and Grandy (1984), we conclude that students' overall perceptions were closer to test feedback than grade feedback, which was good for boys but bad for girls.

Table 14. Ratio of Girl/Boy Ranking on Math HS GPA, Math SAT Scores, and Perceived Math Abilities.

Item	Girl Score Divided by Boy Score
% A+ on Math HS GPA	.85
% A- through A+ on Math HS GPA	.96
% in Highest Group on Math SAT	.62
% in Highest Two Groups on Math SAT	.77
% Estimating Their Math Ability in Top 5%	.52
% Estimating Their Math Ability in Top 10%	.69

Students compare SAT scores at least as avidly as grades. Moreover, students can provide reasons for poor grades -- not doing the homework, not studying. For poor SAT's, students can only supply excuses: "I don't do well on 'standardized' tests," "I don't care about it anyway," "I had a bad day." Thus we suspect that some girls internalize the SAT's under-prediction of their academic performance.

Self-perception and test performance are probably inter-dependent. Table 15 sheds light on this point. It shows the same strong relationship between SAT score and self-perceived ability that previous tables have displayed. In "reading and writing ability," self-perception and SAT scores are similar for both sexes. 49.6% of girls who scored well on the VSAT rank themselves in the top 5%, for instance, compared to only 40.2% of high-scoring boys, but the difference is made up in the next category, top 10%.

In perceived math abilities, however, the sexes behave differently. Girls have a lower perception of their abilities

Table 15. % of Students who Place Themselves In the Listed Percentile Groups in Self-Perceived Abilities, as Affected by SAT Scores.

VSAT Groupings:

	(among girls)				(among boys)			
	Low	Low-Med	Med-High	High	Low	Low-Med	Med-High	High
Self-perceived reading and writing ability								
top 5%	12.1%	15.8%	22.6%	49.6%	8.0%	18.8%	24.0%	40.2%
top 10%	29.1	37.6	39.1	31.2	22.6	36.6	35.4	41.8
top 25%	29.7	24.1	25.6	14.9	38.7	29.7	28.1	13.9
top 50%	21.2	14.3	3.8	0.7	25.5	12.9	9.4	1.6
bottom 50%	4.2	3.0	0.0	0.0	2.9	1.0	1.0	0.8

MSAT Groupings are identical to above:

Self-perceived math ability								
top 5%	2.5%	7.3%	16.2%	38.9%	5.1%	13.1%	21.4%	57.1%
top 10%	9.4	22.7	32.5	36.1	9.0	22.4	41.2	30.7
top 25%	32.5	36.0	35.1	16.7	28.2	34.6	23.7	10.7
top 50%	28.8	25.3	7.1	2.8	42.3	22.4	13.0	1.4
bottom 50%	20.6	6.7	0.6	0.0	11.5	3.7	0.0	0.0

even when controlling for SAT scores. Among high MSAT scorers, for instance, 57% of boys put themselves into the top 5% in math ability, while only 39% of girls did so. Conversely, among low MSAT scorers, 20.6% of girls put themselves in the lower half in math ability, while only 11.5% of males did so. In other words, when the test tells them they are good at math, girls are less likely to believe it.[8]

[8]On the other hand, we found that among girls in the lowest scoring group on the MSAT, 89% say their "ability is higher than the tests indicate," while 81% of the low-scoring boys agree. Thus girls do not simply internalize low MSAT scores.

Do SAT Scores Influence Future Aspirations?

Our students displayed high college aspirations. More than 95% of each sex planned to attend "super-elite," "very strong," or "strong" four-year institutions. High school GPA made a large difference to whether students planned to attend "super-elite" rather than "very strong" colleges. Sex made some independent difference, as Table 16 shows. 52% of A to A+ girls planned to

Table 16. Students Who Plan to Attend Different Types of Colleges, By High school GPA.

Type of College	Students with High school GPAs of			
	A to A+	B+ to A-	B- to B	C+, Lower
Among female students:				
Super-Elite	52.3	18.7	7.4	11.8
Very Strong	30.8	48.4	34.1	11.8
Strong	14.0	30.1	55.7	52.9
Among male students:				
Super-Elite	65.7	36.3	7.7	5.9
Very Strong	22.9	40.5	31.9	20.6
Strong	8.6	20.2	56.6	50.0

attend "super-elite" colleges, compared to 66% of A to A+ boys; among B+ to A- students, the difference is greater. The fact that SAT scores matter most to applicants to competitive "super-elite" institutions implies that the lower SAT scores received by girls with very high grades compared to boys with very high grades might have contributed to girls' lower aspirations.

However, when we looked directly at the influence of SAT scores on college choices, we found that they did not account for the sex difference in super-elite choices. Within each SAT score

category, boys were more likely to attend super-elite colleges than girls. Among high-MSAT girls, for example, 45% plan to attend super-elite colleges; among high-MSAT boys, 51% plan to do so; the difference, 6%, is exactly the same as between all girls

Table 17. Students Who Plan to Attend Different Types of Colleges, By SAT Scores.

Type of College	Students with Verbal SATs				
	Low	Low-Mid	High-Mid	High	All
% of female students choosing:					
Super-Elite	5.5	14.3	19.5	47.5	21.2
Very Strong	29.1	44.4	42.1	39.0	38.1
Strong	58.2	37.6	34.6	11.3	36.4
% of male students choosing:					
Super-Elite	11.7	18.8	26.0	51.6	27.0
Very Strong	27.7	28.7	46.9	31.1	32.9
Strong	51.1	50.5	25.0	13.9	35.5
	Students with Math SATs				
	Low	Low-Mid	High-Mid	High	All
% of female students choosing:					
Super-Elite	5.6	14.7	26.6	45.4	21.2
Very Strong	25.6	39.3	51.3	36.1	38.1
Strong	61.3	42.0	18.2	17.6	36.4
% of male students choosing:					
Super-Elite	6.4	5.6	30.5	51.4	27.0
Very Strong	15.4	40.2	37.4	32.9	32.9
Strong	62.8	51.4	30.5	12.9	35.5

Therefore we cannot lay the difference at the doorstep of the SAT.

VIII. Summary of Major Findings.

We found four important areas of sex-related differences.

First, we found the same under-prediction that other researchers have noted: girls did less well than boys on the

SAT, yet they had higher high-school grades than boys in both English and math.

Second, 17 items were considerably (>10%) easier for one sex, suggesting that ETS's review process doesn't work effectively. (Suggestions for strengthening it follow below.) Specific item content made the greatest difference, rather than whether the item was an analogy, treated geometry, was difficult, etc.

Third, girls' poorer performance was not linked to test anxiety or time pressure. Boys liked math somewhat better and took slightly more math, which explained part but not all of their math SAT lead over girls; liking math adversely affected verbal SAT scores to some extent. Controlling for social class, we still found a score gap favoring boys. Thus social class did not explain the gender gap; we expected it would not, since gender and class aren't systematically tied to each other, and since our students' SES range was constricted. Independently, SES had great impact on SAT scores: children of parents with higher status jobs and more education scored better.

Fourth, when estimating their math and English abilities, both sexes perceived their abilities to be more in line with their test scores than their grades. Unfortunately, this meant that girls perceived themselves to be less able than their grades

would indicate, and less able than boys. Aspirations to "super-elite" colleges behaved similarly.[9]

IX. Implications for Test-Makers.

Our research leads to several conclusions for test-makers.

Reviewers Cannot Detect Biased Content Reliably.

As we understand their procedures, ETS used three procedures for evaluating items during the construction of this test: using deltas to assess the general difficulty of each item, reviewing item content, and calculating item-to-scale (biserial) correlation coefficients.[10] Deltas are used to assemble tests containing the desired number of easy, medium, and difficult questions. Apparently ETS does not use the procedure to investigate different degrees of difficulty among different subgroups of test-takers, so it is not relevant to our topic.

ETS's descriptions of its item review process (Donlon, 1984; ETS, 1987; cf Donlon and Angloff, 1971) do not make clear the details of the process as applied to a given test. Apparently,

---

[9]However, sex differences in these two areas also persisted when SAT scores were controlled for, with men ranking their abilities moderately higher than women and aspiring to "super-elite" colleges at a moderately higher rate.

[10]ETS uses biserial r's, which omit the item's contribution to the overall scale score (Donlon, 1984). We used point-biserial r's for ease of computer programming. The difference in practice is trivial.

proposed items are reviewed to see that they do not offend a minority or sex. Perhaps they are also reviewed to see that they do not obviously favor the subculture and vocabulary of any "subgroup of English speakers" (ETS, 1987). If so, our results call into question the effectiveness of this face validity check. Verbal items with obviously sex-biased content, such as "pendant" and "mercenary," were left on this exam, and they proved to favor one sex or the other by considerable margins.

To assess the effectiveness of ETS's procedure, Loewen replicated it, judging each VSAT item for male- or female-bias, simply on the basis of subject matter, before looking at any results. Loewen predicted that girls would do better on 7 items, boys on 3. (Loewen was not attempting to predict the magnitude of the difference, merely its direction.) Results proved his predictions correct on 9 items, wrong on 1. We were surprised that the ETS review process did not weed out culturally-loaded items that were noted by a single untrained observer, especially since an ETS researcher made similar predictions and achieved similar results more than a quarter century ago, before ETS's review process was in place (Coffman, 1961).

On the math sections, Loewen made no predictions, but 3 of the questions on which boys showed the greatest advantage dealt with boys' camp and basketball team statistics; again, it would seem that item review should have caught and removed such overt bias.

However, we do not recommend that ETS simply hire better reviewers. Our replication of their face validity check left us unconvinced that this was an effective way to detect and remove biased items. Although his judging of items was more effective than ETS's, Loewen missed several on which one sex scored >10% better than the other. The content of one item, "sheen" opposite to "dull finish," obviously drew upon the subculture and vocabulary of girls. But other items on which one sex showed a peculiar advantage weren't so obviously biased in content, particularly on the math test. Our knowledge as to differences in vocabulary and cognitive styles among different racial groups and between boys and girls is modest; hence even after our results flagged an item as favoring one sex or the other, we weren't always able to explain why. Therefore we doubt that sex bias (or racial or class bias) can be predicted consistently on the basis of item content.

Item-to-Scale Correlations Cannot Detect Bias.

After items have been judged fair, or at least inoffensive, ETS then puts them on experimental sections of the SAT and computes item-to-scale r's. Such correlations have no mitigating effect on sex (or racial) bias. Indeed, to the degree that the test as a whole favors affluent, white, or male subcultures, using r to screen items will maintain or increase bias on sex, class, or racial lines.

An example can clarify this point. Imagine a verbal SAT item that tapped working-class culture, such as item 3, "spline is to miter as straw is to mud," from the "Loewen Low-IQ Test." (Loewen, 1979) It involves difficult reasoning and might help predict which students from working-class culture were most capable of that reasoning, but it would never get past the biserial r hurdle, because upper- and middle-class students would get it wrong, while some working-class students would get it right. Since SAT scores are strongly class-related, "spline" would not correlate well with overall scores. Hence no item favoring working-class culture is likely to be included on any SAT. Indeed, we found that point-biserial r's for "classist" items were higher than for class-fair items on this test.

The situation is similar regarding sex and the math SAT: because girls score worse than boys, any item on which girls excelled would be unlikely to have a robust biserial r, so ETS would drop it. Indeed, we note that the 5 most pro-boy items on the math SAT show r's averaging .45, while for the 10 items on which girls approximately equalled boys, average r = .30.[11] The r test probably acts to increase sex bias on the math SAT. On the verbal SAT, using the biserial r to qualify an item has no systematic effect on sex bias, because boys and girls are roughly

---

[11]Seven of these items were easy (>80% of all test takers got them right.). High r's on easy items are hard to achieve, partly because errors unrelated to content -- sloppy marks, using the wrong answer column, and the like -- become an appreciable proportion of all errors, and such random errors act like "noise in the system" to reduce r's. On the 3 other items on which girls excelled, r = .37, modestly lower than the r on the items favoring boys.

equal in numbers and performance. Thus pro-boy and pro-girl verbal items can pass this hurdle and get included. A pro-girl math item would probably not make it onto the test. Neither would a pro-minority item.

New Procedures Are Needed To Avoid Item Bias.

Because ETS procedures proved unable to identify sex-biased items on this SAT, different procedures are needed to reduce test bias. We suggest two. First, ETS should publicize the studies it now conducts on the relationship between SAT scores and first-semester college grades, and should perform more research correlating performance on each SAT item with those grades. Such research would lead to future SATs with higher correlations to first-year college GPA, which would probably reduce test bias and certainly increase test defensibility.

We suspect that some SAT items predict first-year college GPA poorly, because 24 items on this SAT were very poor predictors ($r < .1$) of high school GPA. Since the correlation between an item and college GPA is a more important index of validity than its biserial correlation, items that fail to show a reasonable r should be considered for deletion from the test. We believe that items that are particularly "loaded" toward white, upper-class, and male subcultures will be modestly more likely to fail to meet this criterion.

Second, test-makers need to evaluate items after their inclusion on experimental sections, comparing the percentage correct among each sex (and among races and social classes) to identify extreme items.[12] Mindless deletion of items favoring one sex or ethnic group may not make the SAT fairer, however, for a test might contain no such extreme items, yet still be biased against women or men (or racial or ethnic groups). To put this another way, including male-biased items like "mercenary" might be defensible, providing they are balanced with enough female-biased items like "sheen." A fair test assembly procedure must first be aware of biased items and then use this knowledge to construct fair tests. This brings us to the issue of balance, or overall test fairness.

Constructing Sex-Biased and Sex-Equal Verbal Tests.

The existence of verbal SAT items that markedly favor one sex or the other indicates that the 10 point "gender gap" suffered by girls nationally is manipulable by the content of the included items. Test-makers could easily construct a test on which one sex nationally scored as much as 50 points better than the other. On this test, if we deleted the 10 items that favored boys the most, replacing them with items similar to the 10 items that most favored girls, girls nationally would outperform boys by about 4 points. This change would be accomplished solely with items that could pass through ETS's screening process.

---

[12]We understand ETS may have instituted this procedure within the last year, after the construction of this SAT.

Thus "balance" has primarily a political, not intellectual, definition. ETS has long known that "categories designated 'world of practical affairs' and 'science' are typically easier for males, whereas the categories designated 'aesthetics/ philosophy' and 'human relationships' are easier for females." ETS apparently believes that its changes in the verbal SAT, which substituted a male advantage for the previous female advantage, are "balanced" and "seem to accomplish their purpose" (Donlon, 1984, p. 52).

We do not agree. Since any difference between boys' and girls' means is dependent upon inclusion or exclusion of questions favoring one sex or the other, we do not see how the observed national 10 point difference can somehow be considered "real" or how the test that created this difference can be considered "balanced." As we have seen, items could be included so that no difference in group means for boys and girls would result. We suggest that this be done. More generally, we suggest that as ETS studies the performance of subgroups, items that particularly favor males, whites, and the affluent should be removed or balanced with items favoring females, minorities, and working-class culture.

Constructing Sex-Biased and Sex-Equal Math Tests.

As with the verbal test, we can alter averages for males and females by replacing existing math items favoring boys with items

similar to current items that favored girls. Because boys outscored girls on most math items, a sex-equal math test cannot be constructed solely from existing questions. On the math SAT nationally, boys now outscore girls by about 47 points on ETS's 200-800 scale. Since the difference between boys' and girls' means partly derived from questions favoring males by margins of >10%, at least 3 of which contained overtly pro-boy verbal content, we cannot consider all of this difference "real." If the 10 most pro-male items were replaced with items similar to the 10 most pro-female items, boys nationally would outscore girls by about 29 points. More than a third of the existing math "gap" suffered by girls nationally would be eliminated by excising these 10 items.

Only one math item had any verbal content related to girls, and that consisted solely of the proper noun "Judy" in Section 2 item #11, "Judy doubles k, adds 12..." Otherwise that item too was gender-free. On it girls did rather well, .5% below boys. In contrast, on two items set in a boys' camp, boys outperformed girls by 12.3% and 15.6%. And the largest sex-related difference of all, 27%, occurred on the item dealing with basketball team statistics.

Because the SAT math gap is not replicated in school performance, and because we have found that verbal content of math questions influenced scores by sex, we recommend that ETS revise its math questions to insert verbal content that overtly includes girls' subculture and female names and omits boys'

subculture and male names -- just the reverse of current practice on this SAT (except for "Judy"). We estimate this might lead to a further increment of perhaps 5 points in girls' scores, relative to boys' (cf. Donlon, et al., 1977). Moreover, adding items with female verbal content might create items with pro-girl differences, which the test does not now contain. Several studies have shown that females perform better on questions that refer to females or whose content reflects their cultural experience (Donlon, Ekstrom, and Lockheed, 1979; Dwyer, 1979; Stricker, 1982).

Summary of Policy Recommendations for Sex-Fairer Tests.

1. Remove items from the test that have large response differences between the sexes, unless they are balanced by other items.

2. Since male and female mean scores on the verbal test are arbitrary and manipulable by the test-maker, manipulate them so that males and females score equally. Areas where females excel, such as writing and human relations, either are not evaluated by the SAT or are downplayed in favor of math, science, and business items. The SAT should test a more balanced array of skills, because our society needs to include both sexes equally in its talent search.

3. Since girls have more anxiety about the math SAT than boys, score worse on it, and yet do as well or better in math

courses, maximize girls' sense of comfort on math items by ensuring that their verbal content favors girls, not boys as on this test. (Although it is beyond the scope of our research, we also recommend continued efforts to encourage females to take and excel in mathematics.)

4. Make widely available the validity studies correlating test scores and first-year college performance, so that consumers are aware of the level of their predictive accuracy at various institutions and researchers are aware of this valuable database.

REFERENCES

Adams, Raymond J. 1985. "Sex and Background Factors: Effect on ASAT Scores." Australian Journal of Education, Volume 29, No. 3, pp. 221-30, November.

ACT (American College Testing Program). 1973. Assessing Students on the Way to College. Iowa City: American College Testing Program.

AP (Associated Press). 1987. Drug May Help the Overanxious on S.A.T.'s. New York Times, 10/22, p. A27.

Becker, Betsy Jane. 1983. "Item Characteristics and Sex Differences on the SAT-M for Mathematically-Able Youths." Montreal, Canada: Paper presented at the Annual Meeting of the American Educational Research Association, April.

Bem, S.L., Bem, D.V. 1970. "We're All Nonconscious Sexists." Psychology Today, April.

Behnke, Michael, J. 1987. Washington, D.C.: Testimony Presented to the Congressional Subcommittee on Civil and Constitutional Rights, April 23.

Breland, H.M. 1978. Population Validity and College Entrance Measures. Research and Development Report 78-79, No. 2. New York: College Entrance Examination Board.

Clark, Mary Jo and Grandy, Jerilee. 1984. Sex Differences in the Academic Performance of Scholastic Aptitude Test Takers. College Board Report No. 84-43. New York: College Entrance Examination Board.

Coffman, William E. 1961. "Sex Differences in Responses to Items in an Aptitude Test." Lansing, MI: National Council on Measurement in Education, Eighteenth Yearbook, pp. 117-124.

Cordes, Colleen. 1986. Monitor, American Psychological Association, June.

Donlon, Thomas F. 1973. Content Factors in Sex Differences on Test Questions, Research Memorandum 73-28. Princeton, NJ: ETS.

Donlon, Thomas F. 1977. "Sex Differences in Test Speededness on the Scholastic Aptitude Test." Ellenville, NY: Paper presented at the Annual Meeting of the Northeastern Educational Research Association, October 26-28.

Donlon, Thomas F., and Angoff, Wiliam H. 1971. "The Scholastic Aptitude Test." The College Board Admissions Testing Program, William H. Angoff, Ed., pp. 15-47. New York: College Entrance Examination Board.

Donlon, Thomas F., Ekstrom, Ruth B., and Lockheed, Marlaine E. 1979. "The Consequences of Sex Bias on the Content of Major

Achievement Test Batteries." Measurement and Evaluation in Guidance, Volume 11, No. 4, January.

Dwyer, Carol A. 1976a. "Test Content and Sex Differences in Reading." The Reading Teacher. May.

Dwyer, Carol A. 1976b. "Test Content in Mathematics and Science: The Consideration of Sex." San Francisco: paper presented at annual meeting of American Educational Research Association.

Dwyer, Carol A. 1979. "The Role of Tests and Their Construction in Producing Apparent Sex-Related Differences." Sex Related Differences in Cognitive Functioning, Wittig and Petersen, Eds. New York: Academic Press.

ETS Sensitivity Review Process. 1987. Princeton, NJ: ETS.

Graf, Richard G. and Riddell, Jeanne C. 1972. "Sex Differences in Problem-Solving as a Function of Problem Context." The Journal of Educational Research, Volume 65, No. 10, July-August.

Green, Donald Ross. 1987. "Sex Differences in Item Performance on a Standardized Achievement Battery." New York: Paper presented at the annual meeting, American Psychological Association.

Jacobs, Janis E. and Eccles, Jacquelynne S. 1985. "Gender

Differences in Math Ability: The Impact of Media Reports on Parents." Educational Researcher, March.

Klein, Susan S. 1986. "Why Do Females Receive Higher Course Grades, But Often Lower Standardized Achievement and Aptitude Test Scores Than Males." Washington, DC: Paper presented at National Center for Fair and Open Testing Conference, December.

Linn, Robert L. 1973. "Fair Test Use in Selection." Review of Educational Research, Volume 43, pp. 139-161.

Linn, Robert L. 1978. "Single-Group Validity, Differential Validity, and Differential Prediction." Journal of Applied Psychology, Volume 63, No.4, pp. 507-512.

Linn, Robert L. 1982. "Selection Bias: Multiple Meanings." Presidential Address to the Division of Evaluation and Measurement at the Annual Meeting of the American Psychological Association, August.

Loewen, James W. 1979. "Introductory Sociology: Four Classroom Exercises." Teaching Sociology, Volume 6, No. 3 (April), pp. 221-244.

Maccoby, E.E. and Jacklin, C.M. 1974. The Psychology of Sex Differences. Stanford, CA: Stanford University Press.

McCarthy, K. 1975. "Sex Bias in Tests of Mathematical Aptitude." New York: City University of New York, doctoral dissertation.

McPeek, W. Miles and Wild, Cheryl L. 1987. "Characteristics of Quantitative Items that Function Differently for Men and Women." New York: Paper presented at the Annual Meeting of the American Psychological Association, August.

Milton, G. A. 1957. "The Effects of Sex Role Identification Upon Problem Solving Skill." Journal of Abnormal and Social Psychology, Volume 55, pp. 208-212.

Milton, G. A. 1958. "Five Studies of the Relation Between Sex Role Identification and Achievement in Problem Solving," Technical Report No. 3, Department of Industrial Administration, Department of Psychology. New Haven, CT: Yale University, December.

Novick, M. R. 1982. "Educational Testing: Inferences in Relevant Subpopulations." Educational Researcher, Volume 11, pp. 4-10.

Ramist, Leonard and Arbeiter, Solomon. 1986. Profiles, College-Bound Seniors, 1985. New York: CEEB.

Rosser, Phyllis. 1987. Sex Bias in College Admissions Tests: Why Women Lose Out. Cambridge, MA., National Center for Fair and Open Testing.

Schrader, William B. 1984. Three Studies of SAT-Verbal Types. Report No. 84-7. New York: CEEB.

Schmitt, Alicia P. and Dorans, Neil J. 1987. "Differential Item Functioning for Minority Examinees on the SAT." New York: Paper presented at the Annual Meeting of the American Psychological Association, August.

Strassberg-Rosenberg, B. and Donlon, T. F. 1975. "Content Influences on Sex Differences in Performance on Aptitude Tests." Washington, DC: Paper presented at the Annual Meeting of the National Council on Measurement in Education, March.

Stricker, Lawrence J. 1982. "Identifying Test Items That Perform Differentially in Population Subgroups: A Partial Correlation Index." Applied Psychological Measurement, Volume 6, No. 3, Summer, pp. 261-273.

The College Board. 1987. National Report: 1987 Profile of SAT and Achievement Test Takers. New York: CEEB.

Welsh, W. W., Anderson, R. E., and Harris, L. J. 1982. "The Effects of Schooling on Mathematics Achievement." American Educational Research Journal, Volume 19, pp. 145-153.

Wild, Cheryl L., Durso, Robin, and Rubin, Donald B. 1982. "Effect of Increased Test-Taking Time on Test Scores by Ethnic Group, Years Out of School, and Sex." Journal of Educational

Measurement, Volume 19, No.1, Spring, pp. 19-28.

APPENDIX A: ITEMS WITH EXTREME DIFFERENCES BY SEX

Section 1

1. SETBACK: (A) commotion
(B) variation (C) eagerness
(D) concentration (E) improvement

5. SHEEN: (A) uneven length (B) dull finish
(C) strong flavor (D) narrow margin
(E) simple shape

23. The author's tone can best be described as which of the following?

(A) Whimsical (B) Confidential (C) Narrative
(D) Instructive (E) Speculative

44. MERCENARY:SOLDIER:: (A) censor:author
(B) hack:writer (C) agent:performer
(D) fraud:artist (E) critic:subject

Section 4

21. PENDANT:JEWELRY:: (A) frame:picture
(B) cue:drama (C) violin:music
(D) mobile:sculpture (E) poetry:prose

24. LOVE:REQUITE:: (A) attack:retaliate
(B) proposal:write (C) problem:worry
(D) film:review (E) law:domineer

31. Perrot betrays Wilson by revealing that

(A) Dawson's presence should be no surprise to Wilson
(B) Perrot's wife had expected Wilson's arrival
(C) Wilson has ignored the plight of the victims
(D) Wilson has been involved in a scandal in the city
(E) Wilson has lied about his age

Section 2

8. A certain sprinkler releases water at the rate of 150 liters per hour. If the sprinkler operates for 80 minutes, how many liters of water will be released?

(A) 170 (B) 200 (C) 225
(D) 230 (E) 250

Questions 15-16 refer to the following information.

CAMP SCHEDULE OF CHORES

Order of Assignment	Chore
1	Make beds
2	Mop floors
3	Clean windows
4	Pick up litter
5	Empty waste cans
6	Clean bathrooms
7	Pick up mail
8	Inspect cottage
9	Deliver laundry

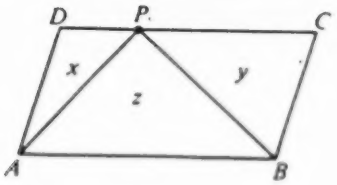
A boys' camp had 200 empty cottages. When 1,800 boys arrived, they were numbered serially starting with 1 and were assigned, in order, to cottages with 9 boys to a cottage. The first 9 boys were assigned to the 1st cottage, the second 9 to the 2nd cottage, and so on. In each cottage, each boy was assigned to chores according to his number, with the boy having the lowest number in each cottage assigned to the first chore, and so on.

15. What chore will the 994th boy have?

(A) Mop floors
(B) Clean windows
(C) Pick up litter
(D) Clean bathrooms
(E) Deliver laundry

16. What was the number of the boy in the 86th cottage whose assignment was to "inspect cottage"?

(A) 766
(B) 773
(C) 774
(D) 775
(E) 782



19. In parallelogram $ABCD$ above, P represents any point on side DC. If x, y, and z are the areas of the three triangles shown, which of the following CANNOT be the ratio of x to y to z?

(A) 1 to 3 to 4
(B) 7 to 8 to 15
(C) 3 to 7 to 10
(D) 4 to 8 to 12
(E) 2 to 5 to 8

20. If $\frac{1}{6}$ is written as a decimal to 200 places, what is the sum of the first 100 digits to the right of the decimal point?

(A) 55
(B) 100
(C) 350
(D) 595
(E) 600

21. A high school basketball team has won 40 percent of its first 15 games. Beginning with the sixteenth game, how many games in a row does the team now have to win in order to have a 55 percent winning record?

(A) 3
(B) 5
(C) 6
(D) 11
(E) 15

22. If $-3 < a < 7$ and if $-2 < b < 0$, which of the following must be true for $(a - b)$?

(A) $-5 < (a - b) < 7$
(B) $-3 < (a - b) < 7$
(C) $-1 < (a - b) < 7$
(D) $-3 < (a - b) < 9$
(E) $-1 < (a - b) < 9$

25. If n is one of three consecutive odd integers, then the possible values of the sum of the 3 integers include which of the following?

I. $3n + 3$
II. $3n$
III. $3n + 6$

(A) I only (B) II only (C) III only
(D) I and III (E) II and III

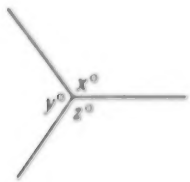
Section 5

COMPARISON QUESTIONS

Answer: A if the quantity in Column A is greater;
B if the quantity in Column B is greater;
C if the two quantities are equal;
D if the relationship cannot be determined

AN E RESPONSE WILL NOT BE SCORED.

	Column A	Column B
	Two of the three sides of a right triangle R have lengths 7 and 10.	
17.	Length of the remaining side of R	10

	Column A	Column B
		
	$-1 < x < 0$	
25.	x^2	$-x$

APPENDIX B: QUESTIONNAIRE

Survey for Research on SAT Tests

This anonymous questionnaire is designed to help researchers uncover problems students encounter on standardized tests. None of this material will go to your school or be used with your name attached. We need your help -- please fill out each question carefully! If you can't answer a question or choose not to, please move on to the next item. Thank you!

1. What is your grade level in school?

   (A) 12th grade or no longer in H.S.
   (B) 11th grade
   (C) 10th grade
   (D) 9th grade or earlier

2. From this list, which is your favorite subject in high school?

   (A) English
   (B) Math
   (C) Social Studies
   (D) Science
   (E) Foreign Language

3. What is your second favorite subject?

   (A) English
   (B) math
   (C) social studies
   (D) science
   (E) foreign lang.

4. How many years of math have you had in high school, from ninth grade
   until now (include this year, if you are taking math this year)?

   (A) six or more
   (B) five
   (C) four
   (D) three
   (E) two or less

5. How many years of English have you had in high school, from ninth grade
until now (include this year, if you are taking English this year)?

(A) six or more
(B) five
(C) four
(D) three
(E) two or less

6. How many years of science have you had in high school, from ninth grade
until now (include this year, if you are taking science this year)?

(A) six or more
(B) five
(C) four
(D) three
(E) two or less

7. What is your overall grade average in your high school English courses?

(A) A to A+ (93-100)
(B) B+ to A- (87-93)
(C) B- to B (80-86)
(D) C to C+ (73-79)
(E) C- or lower (72 or lower)

8. What is your overall grade average in your high school math courses?

(A) A to A+ (93-100)
(B) B+ to A- (87-93)
(C) B- to B (80-86)
(D) C to C+ (73-79)
(E) C- or lower (72 or lower)

9. What is your overall grade average in all your high school courses?

(A) A to A+ (93-100)
(B) B+ to A- (87-93)
(C) B- to B (80-86)
(D) C to C+ (73-79)
(E) C- or lower (72 or lower)

10. Thinking of your entire high school class in grade average, are you in the:

(A) top 5%
(B) top 10%
(C) top 25%
(D) top 50%
(E) bottom 50%

11. How do you think you compare with other people your own age in your reading and writing ability?

(A) top 5%
(B) top 10%
(C) top 25%
(D) top 50%
(E) bottom 50%

12. How do you think you compare with other people your own age in your ability in math?

(A) top 5%
(B) top 10%
(C) top 25%
(D) top 50%
(E) bottom 50%

13. Do you feel your past test scores on standardized tests (PSAT, etc.) are accurate?

(A) No, my ability is higher than the tests indicate.
(B) Yes, they do reflect my ability.
(C) No, my ability is lower than the tests indicate.

14. How do you feel about the SAT?

(A) extremely anxious
(B) moderately anxious
(C) somewhat anxious
(D) not anxious at all.

15. Have you taken any other coaching course before this?

(A) yes, in school
(B) yes, outside of school
(C) no

16. Think about the colleges you plan to apply to. Which of these phrases best describes the kind of college that you realistically plan to attend?

(A) academically "super-elite," such as Ivy League, Bryn Mawr, Cal-tech, Carleton, Chicago, MIT, Stanford, Smith, Swarthmore, Wesleyan, Williams.
(B) academically very strong, such as Bates, Berkeley, Duke, Georgetown, Johns Hopkins, Michigan, Vermont, Virginia, West Point, Wisconsin.
(C) academically strong, such as Fordham, Illinois, North Carolina, Penn State, NYU, Rutgers, SUNY, CUNY, UConn.
(D) academically adequate, such as Monmouth (NJ), CW Post, Pace, Sacred Heart (CT), small state colleges, etc.
(E) do not plan to go to a four-year college.

17. What is your age?

(A) 18 and over
(B) 17
(C) 16
(D) 15
(E) 14 and under

18. Sex: (A) Female
(B) Male

19. Ethnic group:

(A) black (Afro-American)

(B) white (not including Hispanic)
(C) Hispanic (Puerto Rican, Cuban, Mexican-American,etc.)
(D) Asian-American
(E) other (including Native American Indian)

20. What is your father's occupation? (Use these categories as accurately as you can. If he is retired, deceased, or not working, answer for his last job.)

(A) lawyer; MD; architect; college professor; manager or owner of medium to large business; high executive in large company

(B) pharmacist; engineer; veterinarian; manager or owner of small business; lower executive in large company; school teacher; pilot; minister

(C) social worker; insurance; real estate salesman; electrician; Armed Forces; foreman; police

(D) carpenter; industrial worker; clerk; sales clerk; truck driver

(E) janitor; carpenter's helper; laborer.

21. What is your mother's occupation? (Use these categories as accurately as you can. If she is retired, deceased, or not working, answer for her last job.)

(A) lawyer; MD; architect; college professor; manager or owner of medium to large business; high executive in large company

(B) pharmacist; engineer; veterinarian; manager or owner of small business; lower executive in large company; school teacher; pilot; minister

(C) social worker; nurse; insurance; real estate salesperson; electrician; Armed Forces; foreman; police

(D) industrial worker; secretary; sales clerk; cashier; maid; nurse's aide; waitress; seamstress

(E) housewife; mother; volunteer worker; not in paid job at present.

22. What is your father's education? (If you don't know, answer the best you can.)

(A) less than high school graduate
(B) high school graduate
(C) some college
(D) college graduate
(E) graduate or professional (law, M.D., M.A., Ph.D., etc.)

23. What is your mother's education?

(A) less than high school graduate
(B) high school graduate
(C) some college
(D) college graduate
(E) graduate or professional (law, M.D., M.A., Ph.D., etc.)

24. Are you attending a:

(A) public school
(B) parochial school (church-related)
(C) private (prep) school

25. Where is your high school located?

(A) large city (100,000 or more people)
(B) suburb or town in metropolitan area
(C) small city (10,000 to 100,000 people)
(D) rural area or small town (less than 10,000 people, not in metro area)

Thank you again for your help!

Gender Bias in SAT Items, Appendix C: Technical Notes

James W. Loewen

Significance Levels.

Tables 2, 3, 5, 6, 8, and 10 are comparisons of percentages based on sample sizes of approximately 500 (all females compared to all males). On such tables, differences of about 8% are significant at the .01 level; differences of 6% are significant at the .05 level of confidence (two-tailed).

Tables 4, 7, 15, 16, and 17 are comparisons of percentages based on sample sizes of approximately 125 (1/4 of all females, divided into score groups or other groupings, compared to another 1/4, compared to 1/4 of all males, similarly divided, etc.). On such tables, differences of about 17% are significant at the .01 level; differences of about 13% are significant at the .05 level of confidence (two-tailed).

Item "Standardization."

For several years, ETS has been concerned about eliminating what it calls "the contaminating effects of ability differences from the assessment of item fairness." ETS desires to separate out "unexpected differential item performance" from "normal" "differences in subgroup ability." If, for example, we compared sixth-graders to twelfth-graders on the SAT, and sixth-graders did 20% worse than twelfth-graders on, say, item #13, we would want to know how much worse sixth-graders did on all items before concluding that item 13 was biased against sixth-graders. In ETS's terms, we should compare the two groups using some method that does not "exhibit undesirable sensitivities to differences in overall subpopulation ability" (Dorans and Kulick, 1983, pp. 1-3). We will see that ETS simply uses test score as its measure of "overall subpopulation ability."

In recent years ETS has used several statistical techniques to deal with this problem, including the Mantel-Haenszel technique, transformed item difficulty analysis, and a technique it calls "standardization." Standardization has the advantage of being intuitively clear, and ETS seems to be settling upon it as its method of choice. As ETS researchers Dorans and Kulick put it (1983, Abstract), "the primary goal of the standardization approach is to control for differences in subpopulation ability before making comparisons between subpopulation performance on test items."

"Standardization" as used by ETS does not mean what statisticians mean by the term. Statistical standardization means putting a difference between, say, two percentages into standard deviation units. ETS means something quite different; hence we will use quotation marks around the term when using ETS's definition. Dorans and Kulick use female/male differences to illustrate the technique; we will follow their example, using item #44 from the verbal SAT we analyzed, "mercenary is to soldier."

On this item, 48.6% of our girls answered correctly, compared to 64.3% of our boys. Dorans and Kulick would not use that 15.7% difference, however, but would "standardize" by overall scores. To do this, they subtract the % correct among boys who scored 200 on the verbal SAT from the % correct among girls who scored 200 on the verbal SAT; then they do the same for boys and girls who scored 210, and so on, up to those whose overall verbal SAT score was 800. ETS has 61 score groups, from 200 to 800. Then they sum these 61 differences, weighting them by the number of girls in each score category, to calculate d_f, the "standardized" difference.

In practice, this usually results in a percentage difference between the groups which roughly equals the difference between all girls and all boys with which we began, when the two groups have similar overall means. But when the two groups have different means, then "standardization" results in a percentage difference which usually roughly equals the original percentage difference on the item minus the difference in the overall means.[1]

For easier calculation in our example, we grouped our students into 4 "ability" groups rather than 61 and computed d_f, which yielded -15.7%, roughly identical to the raw difference. d_f for other verbal items was similar to the raw differences, as Table 1 shows. This we expected, since our girls scored only .2 worse than our boys overall on the verbal SAT.

Table 1. Raw and "Standardized" Differences on 7 SAT Verbal Items Favoring One Sex by Approximately ≥10%.

Section, Item #, Description	Female % - Male %	d_f
1 #1, "setback," opposite "improvement"	-10.7%	-10.8%
1 #5, "sheen," opposite "dull finish"	+18.3	+21.4
1 #23, author's tone, science passage	-11.8	-11.7
1 #44, "mercenary is to soldier"	-15.7	-15.7
4 #21, "pendant is to jewelry"	+9.6	+10.0
4 #24, "love is to requite"	+14.5	+14.7
4 #31, "betrayal"	+10.2	+10.0

On the math test, "standardization" made a larger difference, as we would expect, since our boys outscored our girls by 3.5 raw points overall. Table 2 compares the raw and "standardized" differences on each item with ≥10% differences.

[1] If the difficulty curve is different for one group, then d_f ≠ the percentage difference minus the mean difference.

Table 2. Raw and "Standardized" Differences on 10 SAT Math Items Favoring One Sex by >10%.

Section, Item #, Description	Female % - Male %	d_f
2 #8, "liters per hour"	-10.3%	-5.7%
2 #15, chore of 994th boy at boys camp	-12.3	-5.5
2 #16, "# of boy with chore at boys camp"	-15.6	-10.9
2 #19, "parallelogram ratios"	-12.2	-5.0
2 #20, "1/6 as decimal, sum of digits"	-10.7	-2.2
2 #21, "basketball team won/loss record"	-27.0	-18.4
2 #22, "<(a-b)<"	-11.0	-4.7
2 #25, "n as odd integer"	-10.8	-2.9
5 #17, "length of right triangle"	-10.7	-3.5
5 #25, "inequalities with X^2, -x"	-10.6	-2.3

By way of contrast, consider the only item on this math SAT with any female verbal content, #11 from section 2, which includes the word "Judy." Boys outperformed girls on this item by 0.5%, making it a relatively good item for girls; when "standardization" is applied, the difference is +5.2%, "favoring" girls. A researcher who used "standardized" differences of $\geq$5% as the criterion to delete items from this math SAT would delete "Judy," while leaving five items on the exam that favor boys by more than 10%!

A terminology problem afflicts ETS's discussions of "standardization." It sounds fine to compare groups matched in "ability" (or in experience, level of schooling, or the like). Good researchers wouldn't compare apples and oranges, or sixth-graders with twelfth-graders. But overall test score is a circular measure of "ability." Consider this passage by Dorans and Kulick:

> Standardization with respect to ability level . . . produces a simple total group comparison, like that based on the overall performance column, which is not confounded by differences in group ability. Standardization accomplishes this goal by using the same standard ability distribution for both groups. (1983, p.4)

A paraphrase could read:

> "Standardization" by total scores produces a simple total group comparison, like that based on the overall performance column, but with the overall group difference removed."

The difference is instructive, because ETS's wording can lure its own researchers into imagining that "standardization" is correct, somehow more scientific, which it is not.

On the contrary, "standardization" can lead to bizarre and paradoxical results. A study of sex differences on the California Achievement Test provides an example (Green, 1987).[2] Of the 72 different forms of the CAT examined, girls outscored boys on 69. Looking at simple percentage differences, girls outscored boys by ≥5% on 1233 of the 3102 different items, while not one item favored boys by ≥5%. But when "standardization" was applied, only 298 of the 3102 items showed differences ≥5%, and most of those items "favored" boys! In other words, if on a given exam girls exceeded boys by 12% overall, yet on a given item girls exceeded boys by "only" 6%, that item would be one of the 1233 on which girls outscored boys by ≥5%, but it would also favor boys by ≥5% after "standardization."

Even when one group performs dramatically worse than another, such as blacks on the SAT, researchers investigating item bias using "standardization" are just as likely to remove items that favor the lower group as items that discriminate against them. Accordingly, "standardization" is not a tool to locate biased items, at least as that term is commonly defined, but instead may mask bias. While "standardization" is an interesting technique and should be used to supplement raw percentage differences, we would suggest examining simple percentage differences, instead.

Scatterplots.

Scatterplots (and correlation and regression) provide another way of analyzing and showing item bias. Figures 1 and 2 are scatterplots for the verbal and math sections of this SAT. Across the x-axis (horizontal) is the % of boys who answered each item correctly; along the y-axis is the % of girls who answered correctly. Each numeral corresponds to the number of items that lie in approximately that location. Correlations are very high, as we would expect: $r = .970$ on the verbal section, .987 on the math. Thus most items lie very close to the regression line.

Nonetheless, and although the plot is small, outlyers can be observed; we have circled those listed on Tables 1 and 2 of this Appendix. On the verbal SAT, the regression equation is

$$y = 3.29 + (.946)x.$$

Heuristically this equation implies that for an item which 0% of males answered correctly, 3.3% of females answered correctly, while for an item which 100% of males answered correctly, 97.9% of females answered correctly (3.29% + 94.6%). Note also that the regession equation on the math scatterplot is

$$y = -12.3\% + (1.103)x,$$

---

[2]Green used a different statistical manipulation but it had the same effect regarding group means.

Figure 1. Girl Percent Correct Vs. Boy Percent, Verbal SAT.

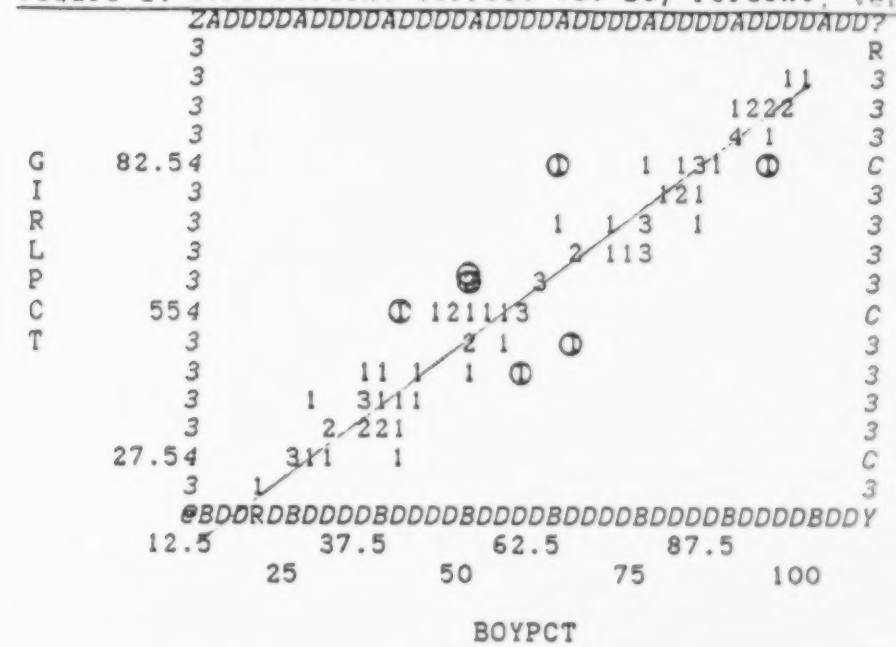
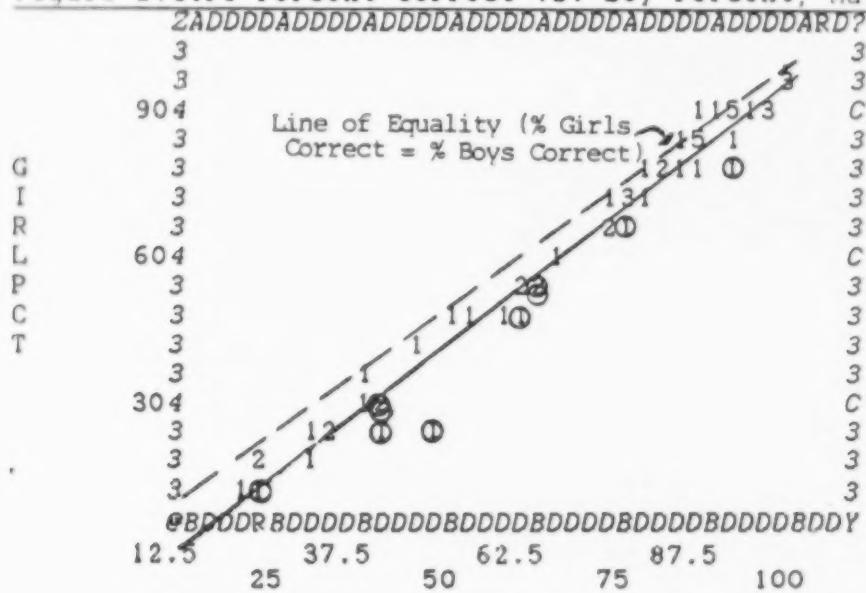


Figure 2. Girl Percent Correct Vs. Boy Percent, Math SAT.



implying that for an item which 0% of males answered correctly, -12.3% of females answered correctly, while for an item which 100% of males answered correctly, 97.0% of females answered correctly. This regression equation restates what we have already observed: that boys outscored girls on the math SAT.

Additional Reference for This Appendix

Dorans, Neil, and Kulick, E. 1983. Assessing Unexpected Differential Item Performance of Female Candidates on SAT and TSWE Forms Administered in December 1977: An Application of the Standardization Approach. Princeton: ETS.